Multiple, correlated covariates associated with differential item functioning (DIF): Accounting for language DIF when education levels differ across languages

Laura E. Gibbons,[1] Paul K. Crane,[1] Kala M. Mehta,[2] Otto Pedraza,[3] Yuxiao Tang,[4] Jennifer J. Manly,[5] Kaavya Narasimhalu,[6] Jeanne Teresi,[7] Richard N. Jones,[8] Dan Mungas[9]

[1]University of Washington; [2]University of California at San Francisco; [3]Mayo Clinic; [4]Program for Appropriate Technology in Health; [5]Columbia University College of Physicians and Surgeons; [6]Duke-NUS Graduate Medical School; [7]Columbia University, New York State Psychiatric Institute, and the Research Division, Hebrew Home for the Aged at Riverdale; [8]Hebrew Senior Life; [9]University of California, Davis, Seattle, WA, USA

Abstract

Differential item functioning (DIF) occurs when a test item has different statistical properties in subgroups, controlling for the underlying ability measured by the test. DIF assessment is necessary when evaluating measurement bias in tests used across different language groups. However, other factors such as educational attainment can differ across language groups, and DIF due to these other factors may also exist. How to conduct DIF analyses in the presence of multiple, correlated factors remains largely unexplored. This study assessed DIF related to Spanish versus English language in a 44-item object naming test. Data come from a community-based sample of 1,755 Spanish- and English-speaking older adults. We compared simultaneous accounting, a new strategy for handling differences in educational attainment across language groups, with existing methods. Compared to other methods, simultaneously accounting for language- and education-related DIF yielded salient differences in some object naming scores, particularly for Spanish speakers with at least 9 years of education. Accounting for factors that vary across language groups can be important when assessing language DIF. The use of simultaneous accounting will be relevant to other cross-cultural studies in cognition and in other fields, including health-related quality of life.

Introduction

There is wide interest in cross-cultural studies of health outcomes. A crucial issue faced in such studies is the psychometric equivalence of test measures across languages and cultures. Test item responses are contingent not only on the underlying ability or trait level (referred to as *ability* here), but also on the language of test administration and the individual's proficiency in that language. Inasmuch as language mediates knowledge structures and schemata,[1] differences in test item responses between individuals with disparate language backgrounds may be associated with differences in language-mediated test behavior, in underlying ability, or both.

Differences in average test scores between groups do not necessarily indicate test bias; these may be valid differences in the ability being measured. Measurement bias is present when individuals from different groups with the same underlying ability have different test scores. An important step in assessing test bias is determining whether test items may have differential item functioning (DIF).[2] DIF is present when individuals who have the same underlying ability from different groups have a different probability of success on an item.[3-5] If item-level bias is present but favors one group for some items and the other group for other items, the overall scale score might still provide an unbiased estimate of true ability. The scale-level impact of DIF is complexly determined by the number of items with DIF, the type of DIF, the pattern of responses, and the direction of DIF. A final determination of test bias is usually made in a qualitative review of the content of any items identified with DIF.

DIF related to language of testing must be accounted for in order to make unbiased comparisons across language groups. However, such studies are rare. Three previous investigations assessed DIF in cognitive tests among Spanish and English speaking elderly people in the United States. Each found DIF related to test language, the first in items from a variety of cognitive screening tests,[6] and the second in the Mini-Mental State Examination.[7] In the third, different methods for detecting DIF were compared using a common Mini-Mental State Examination data set.[8-12] Several other studies identified DIF related to test language but noted that the scale-level impact of DIF was minimal. These studies examined the Cognitive Abilities Screening Instrument (CASI) among Japanese-Americans,[13] the Danish Translation of the SF-36,[14] the Chinese adaptation of the Systemic Lupus Erythematosus Quality of Life Questionnaire,[15] the Fagerstrom Test for Nicotine Dependence,[16] the Alzheimer's Dementia Questionnaire (AD-8),[17] and the Montreal Cognitive Assessment.[18] Some items were found to have DIF related to language among thirteen translations of the European Organisation for Research and Treatment of Cancer (EORTC) QLQ-C30,[19] and among Spanish and English speakers on the Consumer Assessments of Healthcare Providers and Systems.[20] Language is a multifactorial cognitive construct determined by neurobiological, developmental, and psychosocial factors. Among these factors, formal and informal education play key roles in language acquisition and proficiency. Yet, most earlier studies of DIF related to language did not account for differences in educational attainment between language groups.[6-12,14-18,20] An exception is the EORTC QLQ-C30 translation study, where an education term was added to logistic regression models for detecting DIF.[19] The report of DIF in the CASI in Japanese-Americans acknowledged the issue, but noted that it could not be addressed with that sample.[13]

One reason most studies have not evaluated this issue may be that the method for addressing educational differences between language groups is not inherently obvious. There are important technical differences between DIF detection procedures, but each method in essence identifies a group of items that serves as a DIF-free core, anchoring relationships

Correspondence: Laura Gibbons, Department of General Internal Medicine, University of Washington, Box 359780, Harborview Medical Center, 325 Ninth Avenue, Seattle, WA, USA.
Tel. +1.206.744.1842 - Fax. +1.206.744-9917.
Email: GibbonsL@u.washington.edu

Key words: cognitive testing, item response theory, logistic regression, test bias, translation.

Acknowledgments: this research was supported in part by grants NIH AG10220 (Mungas), AG10129 (DeCarli), AG12975 (Haan), K08 AG 022232 (Crane), P50 AG05136 (Raskind), K-01AG025444 (Mehta), AG025308 and AG008812 (Jones), and AG15294 (Lantigua) from the National Institute on Aging.

Contributions: all authors contributed equally to study design, data analysis and interpretation, manuscript final approval.

Conflict of interest: the authors report no conflicts of interest.

Received for publication: 20 September 2010.
Revision received: 31 March 2011.
Accepted for publication: 11 April 2011.

Ageing Research 2011; 3:e4
doi:10.4081/ar.2011.e4

between the test items in the different groups. Most methods for assessing DIF can evaluate only one source of DIF at a time, or require the formation of language/education subgroups. When language groups differ by educational attainment, some language/education subgroups may have sparse data. Methods for identifying anchor items free of DIF related to both language and education have not been previously addressed.

In this study, we sought to compare different strategies for accounting for differences in educational attainment between language groups when evaluating test items for DIF related to language. We evaluated data from an object naming test administered to English- and Spanish-speaking older adults. These adults constitute a relatively large and culturally diverse sample with important differences between English and Spanish speakers in the distribution of the number of years of formal schooling. The object naming test was selected from a larger battery of tests[21,22] because object naming ability represents a language component that is frequently assessed in neuropsychological evaluations, and because the assessment of object naming ability is particularly likely to be affected by DIF. Acquisition of object naming ability occurs over a lifetime and is strongly dependent on experience, making it especially susceptible to differences in linguistic background, education, and cultural experience.

We compared four strategies for assessing DIF related to language: i) simultaneously accounting for DIF related to educational attainment and language in sequential iterative steps; ii) ignoring differences in educational attainment completely; iii) controlling for differences in educational attainment by including it as a covariate in regression models (as in the EORTC study[19]); and iv) ignoring DIF related to educational attainment and language. There are several theoretical reasons we favor the first strategy. DIF related to educational attainment is widely documented in cognitive tests, so ignoring it completely seems unwise. As noted above, scale-level impact is complexly determined by the number of items with DIF, the type of DIF, the pattern of item responses, and the direction of DIF. However, model adjustment applies the same adjustment to everyone in a given subgroup, while incorporating language group-specific item parameters, which seems to ignore item-level variability for education while embracing item-level variability for language. For these theoretical reasons, we favor simultaneously accounting for multiple sources of DIF. Our goal was to compare and contrast the practical implications of these strategies for accounting for DIF related to education and language in an object naming test applied to an ethnically diverse sample.

Materials and Methods

Participants

Participants were recruited between 1998 and 2005 as part of the ongoing development of the Spanish-English Neuropsychological Assessment Scales (SENAS). The SENAS Object Naming Test was administered to 1,755 participants. The vast majority were community-dwelling older adults who were 60 years of age or older at the time of testing (Table 1). Only a quarter (25%) of the Spanish test-takers had more than 6 years of formal schooling, while nearly half (49%) of the English test-takers had more than 12 years of formal schooling (Figure 1). A variety of recruitment methods were used and are described in detail elsewhere.[21] All participants signed informed consent under protocols approved by institutional review boards at University of California at Davis, the Veterans Administration Northern California Health Care System, and San Joaquin General Hospital in Stockton, California.

Materials

The SENAS battery is a multidimensional test battery for the assessment of cognitive functioning in elderly individuals. SENAS scales are new scales that were simultaneously developed with English and Spanish versions. The development process included generation of a large item pool for each scale designed to span a broad range of item difficulty.[21-24] The Object Naming Test assesses the ability to retrieve verbal information from semantic memory stores. Examinees are shown color pictures and asked to identify specific objects within those pictures. A correct response is operationally defined as a word (in either language) that corresponds to the picture being presented. The final selection of 44 items from the larger item pool was based on empirically derived item characteristics. Test items are listed in Table 2.

Statistical methods

We used a hybrid logistic regression-item response theory (IRT) technique to form scores that accounted for DIF (see Appendix). Briefly, we used logistic regression to detect uniform and non-uniform DIF, and then accounted for DIF by using items free of DIF as anchors; items found to have DIF had group-specific item parameters estimated.[25,26] The resulting estimates of object naming ability

Table 1. Participant characteristics by language of test administration [mean (SD) or frequency (%)].

Characteristic	English administration (n=991)	Spanish administration (n=764)	Total (n=1755)
Education (years)	12.6 (3.8)	4.6 (4.3)	9.1 (5.7)
Age (years)	71.7 (7.6)	70.8 (7.4)	71.3 (7.5)
Ethnicity			
Hispanic	379 (38%)	763 (100%)	1142 (65%)
Non-Hispanic Whites	419 (42%)	0 (0%)	419 (24%)
Non-Hispanic African-Americans	172 (17%)	0 (0%)	172 (10%)
Other/missing	21 (2%)	1 (0%)	22 (1%)
Female	526 (53%)	476 (62%)	1002 (57%)
Cognitive status			
Normal	738 (74%)	591 (77%)	1329 (76%)
MCI	173 (17%)	94 (12%)	267 (15%)
Dementia	80 (8%)	79 (10%)	159 (9%)

MCI, mild cognitive impairment.

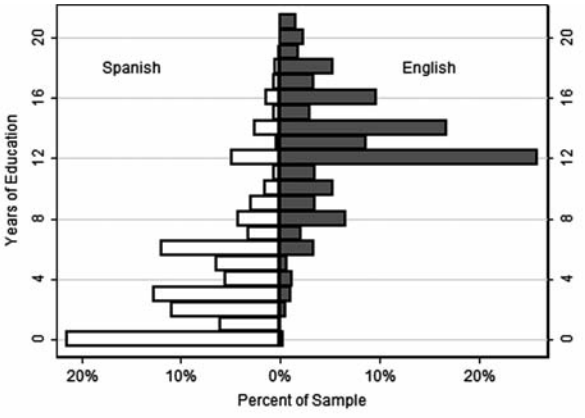

Figure 1. Histogram of years of formal education by language of test administration.

were thus generated from all of the items, but only DIF-free items had the same item parameters for the two languages. Because so few Spanish test-takers had high educational attainment, and so few English test-takers had low educational attainment, our earlier method for accounting for DIF related to more than one covariate[25-28] could not be used here.

We developed a method to address DIF related to language and educational attainment simultaneously (Strategy 1). The goal of the initial steps was to identify anchor items that appeared to be free of DIF related to educational attainment separately in each language group, and that appeared to be free of DIF related to language separately in subgroups defined by educational attainment. Items found to be free of both types of DIF were then used as anchor items for ability assessment. Parameters for the other items were estimated separately for each language-education subgroup. Thus the final run simultaneously accounted for DIF related to language and educational attainment.

In all, we conducted eleven sets of DIF analyses in our initial search for anchor items (Table 3). Items that had no DIF in any of these 11 analyses served as anchors for co-calibrating the scales across language and education groups. We used six language and education specific subgroups for items found to have DIF in at least one of the eleven analyses. The Spanish test-takers were divided into 0-2, 3-7, and 8-19 years of education, and the English test-takers into 0-9, 10-12, and 13-21 years, when computing IRT scores, to reflect the distributions of education in the two language groups. We formed scores accounting for DIF using these six language - education subgroups.

The scores from the simultaneous strategy (Strategy 1) were compared to ignoring education completely while examining items for DIF related to language (Strategy 2), model-level adjustment for educational attainment when accounting for language DIF (Strategy 3),[19] and an unmodified IRT score (Strategy 4). In Strategy 3, we modified our usual DIF detection procedures by adding a term for years of education to all of the logistic regression models. We treated years of education as a continuous, centered covariate in the logistic models for language DIF assessment. For Strategies 1-3, we identified which items had DIF, and who was favored by the DIF. We standardized all scores to have a mean of 100 and a standard deviation of 15. Because all scores have a mean of 100, the overall mean differences between any two scores will be zero. However, the distribution of scores by language and educational attainment may change when accounting for DIF. We tested this by comparing each scoring method to the others, in turn. For each comparison, we calculated the difference between the two scores, and regressed a categorical variable representing the six language - education subgroups on the difference between the two scores.

We needed a reference for comparing changes in an individual's score due to accounting for DIF. In IRT, we recognize that not all scores are estimated with the same precision, and produce a standard error of measurement for each score. Since minimally clinically important differences[29] in scores have not been specified for the object naming test, we used the median standard error of measurement of the unadjusted IRT scores as our threshold.[30] For each person, we took the difference between their original score and their score accounting for DIF. If the absolute difference was larger than the median measurement error, it was identified as a salient difference related to DIF.[30]

Table 2. Non-uniform and uniform differential item functioning in the SENAS object naming test.

Item	Strategy 1: Simultaneous language and education	Strategy 2: Language only	Strategy 3: Language, model-adjusting for education
Mouth-Boca	NU_S[a]	NU_S	
Key-Llave		U_S	
Head-Cabeza	U	U_S	
Suit-Traje			
Hair-Pelo			
Cloud-Nube			
Kitchen-Cocina			
Church-Iglesia			
Pick-Pico	U_S	U_S	
Bird-Ave			
Coin-Moneda	U_S	U_S	
Avocado-Aguacate	U_S	U_S	
Gate-Puerta			
Cemetary-Cementerio			
Lantern-Linterna		U_E	
Knot-Nudo			
Teepee-Tipi			
Spear-Lanza			
Artichoke-Alcachofa			
Llama-Llama		NU_s	NU_s
Castle-Castillo			
Porcupine-Puercoespin		U_S	
Olive-Oliva			
Shrimp-Camarón	U_S	U_S	U_S
Plum-Ciruela	U_S	U_S	
Lobster-Langosta			
Dragonfly-Dragón Volador	U	NU	NU
Mule-Mula	U_S	U_S	
Date-Dátil			
Pheasant-Faisán			
Jewel-Alhaja			
Stone-Piedra		NU_s	NU_s
Fog-Niebla			
Dove-Paloma	U	U_S	
Tapestry-Tapiz	U_E	U_E	
Piñata-Piñata	NU	*na*[b]	
Falcon-Halcón			
Cylinder-Cilindro			
Gable-Gablete			
Parallelogram-Paralelogramo			
Oasis-Oasis			
Scallop-Concha de Peregrino		U_S	
Oyster-Ostra	*na*	*na*	
Cellar-Sótano	*na*	*na*	

[a]S or E subscripts indicate which group was more likely to answer correctly, Spanish or English test-takers, controlling for ability. Direction of DIF varied for most NU DIF. U DIF is not reported if there was NU DIF. [b]na indicates items do not have enough discordance for DIF analysis in the subgroups. NU, Non-uniform, U, uniform.

Secondary analyses

We analyzed DIF related to age and gender. Next, we repeated analyses omitting Hispanics tested in English; there were not enough participants to treat them as a third group. We did this to mitigate potential confounding due to ethnicity and native language.

Results

The three DIF-detection strategies identified different items with DIF (Table 2). Simultaneous accounting for language and educational attainment identified 21 items with DIF. Accounting for language only, while ignoring educational differences, identified 17 items with DIF. Model-level adjustments for differences in educational attainment identified 4 items with DIF.

Mean object naming scores for Spanish and English test-takers, subdivided by educational attainment, are shown in Table 4. English speakers and those with more education had higher scores, even after accounting for DIF related to language and educational attainment ($P<0.001$).

Mean scores may obscure important changes in scores for individuals. To examine the impact on individuals, we compared simultaneous accounting with the other strategies, and found salient scale-level differential functioning related to language for some of the participants (Table 5). The largest changes in score due to strategy choice were observed for the Spanish test-takers with the greatest number of years of formal schooling. The proportion of participants with salient differences with the simultaneous accounting strategy ranged from 2.4-7.9%, depending on the comparison, with changes in both directions. Overall, about 2% of the participants had salient differences when comparing the simultaneous strategy with the other strategies.

Secondary sensitivity analyses confirmed the integrity of the results above. There was no item found with DIF related to age or gender. Results from the analyses omitting Hispanics tested in English were similar to those found in the whole sample.

Discussion

Measurement bias is an important and largely unstudied issue that can affect the interpretation of assessments conducted in diverse populations. This paper addresses problems that have received minimal attention in the literature: the importance of and techniques to account for multiple sources of DIF (such as language and education), especially when individuals grouped by one covariate (such as language) have very different distributions of another covariate known to be associated with DIF (such as education).

This study is an important demonstration of methodology, but also provides information about both the presence of DIF in a specific clinical instrument and its impact on estimation of ability. The overall goal of the SENAS project was a valid assessment of the cognitive functioning of Spanish- and English-speaking individuals. The present paper addresses assessment of one specific cognitive domain, object naming ability, and compares different techniques to use in an attempt to obtain valid cross-cultural inferences from object naming item data collected from English and Spanish speaking study participants. There is extensive literature that differences in educational attainment can lead to DIF in the assessment of cognition, i.e., that heterogeneity in educational backgrounds can interfere with the valid measurement of cognitive functioning.[6,9,31-34] In the data set analyzed here, the educational attainment of Spanish speakers was on average much less than the educational attainment of English speakers (see Figure 1). We therefore were concerned that attempts to address DIF related to language of test administration that ignored educational attainment could lead to incorrect inference.

The three strategies that were compared in this paper produced qualitatively different results. The simultaneous strategy (Strategy 1) incorporated a vigorous search within language groups for DIF related to education, and within education strata for DIF related to language (Table 3). Items that emerged from all eleven sets of preliminary analyses as free of DIF were then used as anchor items for calibrating the scales across language groups. In the final analysis 21 items were identified with DIF. Ignoring educational differences completely (Strategy 2) identified 17 items with

Table 3. The eleven sets of preliminary differential item functioning analyses used to identify anchor items for the item response theory score simultaneously accounting for differential item functioning related to language and educational attainment (Strategy 1).

Type of DIF	Subgroups
Language	High education (≥9 years)
	High education (>median for each language)
	Low education (<9 years)
	Low education (< median for each language)
Education	Spanish test-takers, education <9 versus ≥9 years
	Spanish test-takers, education 0-3 versus ≥4 (median)
	Spanish test-takers, education 0-6, 7-9, and ≥10
	Spanish test-takers, education as a continuous covariate for DIF and 0-2, 3-7, and 8-19 years for IRT analyses
	English test-takers, education <9 versus ≥9 years
	English test-takers, education 0-12 versus ≥13 (median)
	English test-takers, education as a continuous covariate for DIF and 0-9, 10-12, and 13-21 years for IRT analyses

DIF, differential item functioning, IRT, item response theory

Table 4. Means and standard deviations for the object naming scores, by test language and years of education.* The item response theory scores have an overall mean of 100 and SD of 15.

	English administration				Spanish administration			
	0-8 years		9+ years		0-8 years		9+ years	
Score	M	SD	M	SD	M	SD	M	SD
Strategy 1: IRT score Simultaneously accounting for language and education	99.0	10.4	108.7	11.9	88.6	12.3	100.5	10.7
Strategy 2: IRT score accounting for language only	98.9	10.0	109.1	11.8	88.3	12.0	99.7	9.9
Strategy 3: IRT score accounting for language, model-adjusting for education	98.9	10.2	109.1	11.9	88.3	11.9	99.9	10.4
Strategy 4: unmodified IRT score	98.5	10.2	109.0	11.9	88.4	12.0	100.3	10.4
Total score	21.6	5.7	27.4	6.9	16.3	6.6	22.7	6.1
N	152		839		637		127	

* In this table, both language groups are categorized as 0-8 and 9+ for ease of comparison. In Strategy 1 the Spanish test-takers were categorized 0-2, 3-7, and 8-19 years of education, the English 0-9, 10-12, and 13-21 years.

DIF related to language. Despite the similar number of items identified with DIF, DIF impact was qualitatively different between these strategies. This is not surprising, since simultaneously accounting for DIF related to both education and language would be expected to produce qualitatively different results than accounting for DIF related to language alone.

Model-level adjustment for educational attainment by inserting a term in the logistic regression models (uniform adjustment across language groups; Strategy 3) resulted in 4 items with DIF. In terms of DIF impact, model-level adjustment did not produce markedly different results from ignoring DIF related to education. A recent paper employed model-level adjustment to account for several factors between countries in an evaluation of the EORTC QLQ-C30.[19] Model-level adjustment for education did not have much impact in the present study because education DIF was not unidirectional, and, as we found in our analyses for the simultaneous strategy, the effects of education DIF varied by language group.[34]

Group mean object naming scores were similar across all of the strategies analyzed here. Compared to the other strategies, group means from the simultaneous strategy (Strategy 1) were lower in English test takers with more education and slightly higher in Spanish test takers, but differences were small. While the three methods identified 4 to 21 items with DIF, not all the DIF favored the same language/education group, so some of the item-level DIF effect was canceled out at the test level. It may also be that our DIF criteria were too sensitive, picking up differences too small to affect group means.

While mean scores were minimally affected by choice of strategy, it is possible that the choice of strategy may have clinical implications for some individuals. For instance, the simultaneous strategy produced salient differences in scores for 2% of participants. The highest proportion of changes occurred among Spanish test-takers with 9 or more years of education, where up to 8% of the participants were affected. Overall, the most appropriate alternative to a DIF-free test is test scored to account for DIF.

The preliminary analyses for the simultaneous strategy (Strategy 1) revealed that the effects of education DIF were not the same in Spanish and English. Interacting sources of DIF has not been previously recognized as a problem in the literature. A possible exception to this is the work of Jones[35] who used multiple group structural equation modeling in a very large sample. The ideal way of dealing with such DIF is unknown. We think our approach with the simultaneous strategy is reasonable, as it conservatively rejected items as potential anchors if they were found to have DIF in any of the 11 preliminary DIF analyses. These subgroup analyses use smaller sample sizes for anchor item selection, so it is possible that some items with DIF may have been missed, but we think it more likely we were overly thorough. We may be erring on the side of caution when declaring about half the items to not be anchor items, but in the present instance of a reasonable test length with 44 items, this does not seem to be a big problem.

A limitation to this study was the need to categorize education. While the logistic regression framework we used for DIF detection can incorporate continuous covariates (such as years of education), IRT programs require categorical groups to be identified to determine demographic group-specific item parameters when DIF was found. In this regard, we are particularly concerned with two of the preliminary analyses in the simultaneous strategy, in which we dichotomized educational attainment at the median. The median number of years of formal schooling was 3 years for Spanish language test-takers and 12 years for English language test-takers. The small numbers of Spanish speakers with many years of education, and especially the small number of English speakers with few years of education, made meaningful comparisons of individuals with similar educational attainment across the languages difficult (see Figure 1). By making this pragmatic choice, a

Table 5. Differences between object naming scores simultaneously accounting for language and education (Strategy 1), and accounting for language only (Strategy 2), accounting for language with model-adjustment for education* (Strategy 3), and scores that were unmodified (Strategy 4).

Test language and education	N	Strategy 1 minus Strategy 2 (Accounting for language only) Mean[a] (SD)	Range	Salient[b] increase	Salient decrease
			English		
0-8 years	152	0.1 (1.0)	-2.1–3.8	0.0%	0.0%
9+ years	839	-0.4 (1.1)	-8.0–9.1	0.5%	0.4%
			Spanish		
0-8 years	637	0.3 (1.0)	-3.6-9.4	0.3%	0.0%
9+ years	127	0.8 (1.7)	-2.4-11.0	2.4%	0.0%
		Strategy 1 minus Strategy 3 (Accounting for language, model-adjusting for education)			
			English		
0-8 years	152	0.2 (1.2)	-3.6–5.8	0.7%	0.0%
9+ years	839	-0.3 (1.4)	-8.1–9.1	1.0%	0.4%
			Spanish		
0-8 years	637	0.4 (1.6)	-7.6–11.5	0.3%	0.0%
9+ years	127	0.6 (1.9)	-7.8–8.2	3.9%	3.9%
		Strategy 1 minus Strategy 4 (Unmodified)			
			English		
0-8 years	152	0.5 (1.1)	-3.9–6.1	2.0%	0.7%
9+ years	839	-0.2 (1.5)	-8.3–9.1	1.1%	0.5%
			Spanish		
0-8 years	637	0.2 (1.0)	-4.9–7.0	0.3%	0.%
9+ years	127	0.2 (2.0)	-8.5–8.7	0.8%	4.0%

*In this table, both language groups are categorized as 0-8 and 9+ for ease of comparison. In Strategy 1 the Spanish test-takers were categorized 0-2, 3-7, and 8-19 years of education, the English 0-9, 10-12, and 13-21 years. [a]A positive difference indicates that score was greater with simultaneous adjustment. [b]Salient increases or decreases are those greater than the median standard error of the score (see Methods).

person with five years of formal schooling would be characterized as having high educational attainment if they spoke Spanish but low educational attainment if they spoke English. We tried to mitigate categorization problems by identifying DIF-free anchors using many criteria (Table 2). In the present instance, it is likely worse to retain anchor items that actually have DIF than to falsely identify an item as having DIF, since items that have DIF are still used to generate each person's score.[36]

Though we considered educational attainment both as a continuous variable and in a variety of categorizations, any measure of education based on years of formal schooling will not capture differences in the quality of education.[37] This issue is especially relevant in the current study, where Spanish-speaking elderly immigrants and English speaking elders were not educated in similar school systems, the opportunities for attending formal school were discrepant, and the resources available to teachers and schools were drastically different. These factors not only relate to quality of education across language/cultural groups, but also within language/cultural groups. Furthermore, differences in levels of acculturation and in cultural experience that are not fully captured by language used and years of formal schooling may affect test performance, and those differences have not been taken into account. A similar issue is that the English test-takers represent an ethnically diverse group. It is reassuring that secondary analyses omitting Hispanics tested in English resulted in findings similar to the primary analyses.

Simultaneous assessment of DIF is a complicated process, requiring decisions on how to form subgroups and which subgroup analyses to run. This raises the questions of when Strategy 1 is necessary, and when it is feasible. If education is distributed similarly in the two language groups, and years of education mean roughly the same thing in the two language groups, an easier approach is to account for DIF due education using scores that account for DIF due to language.[25-28] It may also be worthwhile to set up interaction terms and look at both education and language in one set of logistic regression models (an extension of Strategy 2). However, in samples such as the one in the current study, where educational levels are so different (Figure 1) and years of education may not be comparable between the two language groups, we think Strategy 1 is the safest option. On the question of feasibility, we cannot make any blanket statements about the number of items needed or the sample sizes required, because both would depend on the distribution of the items with respect to the abilities of the test-takers, and the number of items with DIF. In terms of model convergence, Strategy 1 would require a larger sample than Strategies 2 or 3 because we are dealing with language-education subgroups, rather than just language groups. But we believe that the use of language-education subgroups is necessary to adequately account for DIF in samples like this one. Another disadvantage to Strategy 1 is the number of analyses required; fortunately software exists to facilitate the procedure (see Appendix).

While we compared three different strategies for dealing with DIF in a second covariate that differed across language groups, we did so in the context of a single technique for analyzing items for DIF. There are many methods for detecting DIF[8-12,35] and assessing the impact of DIF.[14,25,26,38-40] All methods of DIF detection require the identification of a set of core items that are free of DIF to serve to anchor the scales. In a test administered in two languages (such as English and Spanish), the actual responses are usually different specific words (e.g., plum in English vs. ciruela in Spanish). The underlying ability tested by this item is the retrieval of a specific word for the pictured fruit from lexical and memory stores. This task is the same in English and Spanish, even though different specific words are obtained. With great care (as outlined above), we think the technique we outlined provides a reasonable way to identify a DIF-free core of items that can serve to anchor the object naming scale in English and in Spanish, improving the validity of inference in cross-cultural analyses of data from this test.

We have compared the simultaneous strategy (Strategy 1) with other possible strategies for addressing language DIF when there is also DIF related to education, and educational attainment differs in the language groups. Ignoring education DIF (Strategy 2) and including a term for educational attainment in logistic regression models (Strategy 3) produced similar results to each other, while the simultaneous strategy - rigorously evaluating items for DIF related to educational attainment in language subgroups, and for language in education subgroups, in an effort to exclude items with DIF from the list of potential anchors - produced qualitatively different results than the other strategies. Further work with simulation studies where true relationships are known may be useful to increase our understanding.

The present study focused on educational attainment and language groups, using one specific cognitive test. This is one example of a more general problem, the evaluation of items for DIF in groups that differ with respect to a second covariate that also is associated with DIF. DIF assessment is a complicated but necessary task when evaluating measurement bias in tests used among ethnically and linguistically diverse populations. As studies enroll increasingly heterogeneous populations, we will need DIF techniques that are up to the task.

References

1. Dabrowska E. Language, mind and brain: Some psychological and neurological constraints on theories of grammar. Washington, DC: Georgetown University Press; 2004.
2. Petersen MA, Groenvold M, Bjorner JB, et al. Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. Qual Life Res 2003;12:373-85.
3. Camilli G, Shepard LA. Methods for identifying biased test items. Jaeger RM, editor. Thousand Oaks: Sage;1994.
4. Holland PW, Wainer H, editors. Differential item functioning. Hillsdale, N.J.: Erlbaum; 1993.
5. Millsap RE, Everson HT. Methodology review: statistical approaches for assessing measurement bias. Appl Psych Meas 1993;17:297-334.
6. Teresi JA, Golden RR, Cross P, et al. Item bias in cognitive screening measures: comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. J Clin Epidemiol 1995;48:473-83.
7. Marshall SC, Mungas D, Weldon M, et al. Differential item functioning in the Mini-Mental State Examination in English- and Spanish-speaking older adults. Psychol Aging 1997;12:718-25.
8. Jones RN. Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: detecting differential item functioning using MIMIC modeling. Med Care 2006;44:S124-33.
9. Crane PK, Gibbons LE, Jolley L, van Belle G. Differential Item Functioning Analysis With Ordinal Logistic Regression Techniques: DIFdetect and difwithpar. Med Care 2006;44:S115-23.
10. Dorans NJ, Kulick E. Differential item functioning on the mini-mental state examination: an application of the Mantel-Haenszel and standardization procedures. Med Care 2006;44:S107-14.
11. Edelen MO, Thissen D, Teresi JA, et al. Identification of Differential Item Functioning Using Item Response Theory and the Likelihood-Based Model Comparison Approach: Application to the Mini-Mental State Examination. Med Care 2006;44:S134-42.
12. Morales LS, Flowers C, Gutierrez P, et al. Item and Scale Differential Functioning of the Mini-Mental State Exam Assessed

Using the Differential Item and Test Functioning (DFIT) Framework. Med Care 2006;44:S143-51.
13. Gibbons LE, McCurry S, Rhoads K, et al. Japanese-English language equivalence of the Cognitive Abilities Screening Instrument among Japanese-Americans. Int Psychogeriatr 2009;21:129-37.
14. Bjorner JB, Kreiner S, Ware JE, et al. Differential item functioning in the Danish translation of the SF-36. J Clin Epidemiol 1998;51:1189-202.
15. Kong KO, Ho HJ, Howe HS, et al. Cross-cultural adaptation of the Systemic Lupus Erythematosus Quality of Life Questionnaire into Chinese. Arthritis Rheum 2007;57:980-5.
16. Yamada H, Acton GS, Tsoh JY. Differential item functioning of the English and Chinese versions of the Fagerstrom Test for Nicotine Dependence. Addict Behav 2009;34:125-33.
17. Koski L, Xie H, Konsztowicz S, Tetteh R. French-English cross-linguistic comparison and diagnostic impact of the AD-8 dementia screening questionnaire in a geriatric assessment clinic. Dement Geriatr Cogn Disord 2010;29:265-74.
18. Koski L, Xie H, Finch L. Measuring cognition in a geriatric outpatient clinic: Rasch analysis of the Montreal Cognitive Assessment. J Geriatr Psychiatry Neurol. 2009;22:151-60.
19. Scott NW, Fayers PM, Bottomley A, et al. Comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. Qual Life Res 2006;15:1103-15.
20. Setodji CM, Reise SP, Morales LS, et al. Differential Item Functioning by Survey Language Among Older Hispanics Enrolled in Medicare Managed Care: A New Method for Anchor Item Selection. Med Care 2011;Mar 18. [Epub ahead of print]
21. Mungas D, Reed BR, Crane PK, et al. Spanish and English neuropsychological assessment scales (SENAS): further development and psychometric characteristics. Psychol Assess 2004;16:347-59.
22. Mungas D, Reed BR, Haan MN, Gonzalez H. Spanish and English neuropsychological assessment scales: relationship to demographics, language, cognition, and independent function. Neuropsychology 2005;19:466-75.
23. Mungas D, Reed BR, Marshall SC, Gonzalez HM. Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. Neuropsychology 2000;14:209-23.
24. Mungas D. Neuropsychological assessment of Hispanics elders: Challenges and psychometric approaches. In: Yeo G, Gallagher-Thompson D, editors. Ethicity and the Dementias Second Edition Washington, D.C.: Routledge; 2006.
25. Crane PK, Gibbons LE, Narasimhalu K, et al. Rapid detection of differential item functioning in assessments of health-related quality of life: The Functional Assessment of Cancer Therapy. Qual Life Res 2007;16:101-14.
26. Crane PK, Cetin K, Cook KF, et al. Differential item functioning impact in a modified version of the Roland-Morris Disability Questionnaire. Qual Life Res 2007;16:981-90.
27. Crane PK, Hart DL, Gibbons LE, Cook KF. A 37-item shoulder functional status item pool had negligible differential item functioning. J Clin Epidemiol 2006;59:478-84.
28. Crane PK, Gibbons LE, Willig JH, et al. Measuring depression and depressive symptoms in HIV-infected patients as part of routine clinical care using the 9-item patient health questionniare (PHQ-9). AIDS Care 2010;22:874-85.
29. Guyatt GH, Osoba D, Wu AW, et al. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002;77:371-83.
30. Crane PK, Narasimhalu K, Gibbons LE, et al. Composite scores for executive function items: demographic heterogeneity and relationships with quantitative magnetic resonance imaging. J Int Neuropsychol Soc 2008;14:746-59.
31. Jones RN, Gallo JJ. Education and sex differences in the Mini-Mental State Examination: effects of differential item functioning. J Gerontol B Psychol Sci Soc Sci 2002;57B:P548-58.
32. Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. Stat Med 2000;19:1651-83.
33. Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: differential item functioning in the CASI. Stat Med 2004;23:241-56.
34. Crane PK. Commentary on comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. Qual Life Res 2006;15:1117-8.
35. Jones RN. Racial bias in the assessment of cognitive functioning of older adults. Aging Ment Health 2003;7:83-102.
36. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin 1993;114:552-66.
37. Manly JJ, Jacobs DM, Touradji P, et al. Reading level attenuates differences in neuropsychological test performance between African American and White elders. J Int Neuropsychol Soc 2002;8:341-8.
38. Crane PK, Cetin K, Cook KF, et al. Differential item functioning impact in a modified version of the Roland-Morris Disability Questionnaire. Qual Life Res 2007;16:981-90.
39. Teresi JA. Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics. Med Care. 2006;44:S152-70.
40. Zumbo BD. A handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999.
41. Muraki E, Bock D. PARSCALE for Windows. 4.1 ed. Chicago: Scientific Software International; 2003.
42. Lord FM, Novick MR. Statistical theories of mental test scores, with contributions by Allan Birnbaum. Reading, MA: Addison-Wesley; 1968.
43. Muthén LK, Muthén BO. Mplus: statistical analysis with latent variables. Fifth ed. Los Angeles, CA: Muthén & Muthén; 1998-2007.
44. Crane PK, Gibbons LE, Jolley L, van Belle G, Selleri R, Dalmonte E, et al. Differential item functioning related to education and age in the Italian version of the Mini-mental State Examination. Int Psychogeriatr 2006;18:505-15.
45. StataCorp. Stata Statistical Software: release 10. College Station, TX: StataCorp LP; 2007.
46. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. Am J Epidemiol 1993;138:923-36.